AFRL-OSR-VA-TR-2013-0116

Hierarchical Kernel Machines:

The Mathematics of Learning Inspired by Visual CortexTitle, i.e, Synthetic Aperture Ladar for Tactical Imaging

Tomaso Poggio, Stephen Smale

Massachusetts Institute of Technology

March 2013
Final Report

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
14-02-2013	Final Report	08/15/2009 - 08/14/2012

4. TITLE AND SUBTITLE

Hierarchical Kernel Machines:
The Mathematics of Learning Inspired by Visual Cortex

5a. CONTRACT NUMBER

FA9550-09-1-0606

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Poggio, Tomaso
Smale, Stephen

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge MA 02139

8. PERFORMING ORGANIZATION REPORT NUMBER

Cage: 80230
DUNS:001425594

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

USAF, AFRL DUNS 143574726
AF Office of Scientific Research
875 North Randolph St., RM 3112
Arlington VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-OSR-VA-TR-2013-0116

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A: Approved for public release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Understanding how the brain works and reproducing its central capabilities in computers is arguably one of the greatest problems in science and engineering. This project directly contributes to this challenge from both a mathematical and an applied point of view. In particular, we have developed a mathematical description of a family of hierarchical architectures for learning, comprised of a collection of definitions, lemmas and theorems which collectively highlight important and salient properties of such architectures. Most important among these properties is the notion of invariance. The theory we have developed characterizes how and why a hierarchical architecture can offer better generalization from few examples in terms capturing and exploiting symmetries in the physical world by way of learning invariances. A comprehensive suite of distributed, GPU-enabled software tools was developed to quickly test hypotheses and validate the theory on large-scale, real-world datasets.

15. SUBJECT TERMS

machine learning, computer vision, hierarchical models, kernel machines

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Tomaso Poggio
U	U	U	UU	11	19b. TELEPHONE NUMBER *(Include area code)* 617-253-5230

Reset

[Final Performance Report]

Hierarchical Kernel Machines: The Mathematics of Learning Inspired by Visual Cortex

Grant Award No: FA9550-09-1-0606
Program Manager: Dr. Jay Myung

Principal Investigators

Tomaso Poggio
Dept. of Brain and Cognitive Sciences
Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology

Stephen Smale
City University of Hong Kong
Visiting Professor, MIT

Submission Date: February 14, 2013

Abstract

Understanding how the brain works and reproducing its central capabilities in computers is arguably one of the greatest problems in science and engineering. This project directly contributes to this challenge from both a mathematical and an applied point of view. In particular, we have developed a mathematical description of a family of hierarchical architectures for learning, comprised of a collection of definitions, lemmas and theorems which collectively highlight important and salient properties of such architectures. Most important among these properties is the notion of invariance. The theory we have developed characterizes how and why a hierarchical architecture can offer better generalization from few examples in terms capturing and exploiting symmetries in the physical world by way of learning invariances. A comprehensive suite of distributed, GPU-enabled software tools was developed to quickly test hypotheses and validate the theory on large-scale, real-world datasets.

Since the last interim report was filed, we have finalized the software tools, completed our empirical evaluation and finalized the theoretical development. In particular, we empirically investigated learning data representations at each layer based on PCA applied to image patches drawn from both static image sets and video sequences. We conducted an expanded set of simulations studying further invariance properties, including learned invariance to both affine transformations and nonlinear transformations. A theoretical analysis involving group-theoretic tools was completed. We carried out a theoretical analysis of statistical learning of data representations more generally (and abstractly), and suggested algorithms for coding data at each layer of a hierarchy, as well as for classifying the output of a hierarchical representation when there are many categories, themselves enjoying hierarchical structure. Finally, we pursued an application of hierarchical kernels to a peptide binding problem arising in the design of vaccines. This involved the development and implementation of new kernels defined on spaces of strings describing amino acid sequences.

1 Introduction

When compared with the learning abilities of biological organisms, modern learning theory - such as the theory of kernel machines - is confronted with two intriguing challenges. The first is the poverty of stimulus problem: organisms can learn complex tasks from far fewer examples than our present learning theory and learning algorithms predict. For instance, discriminative algorithms such as SVMs can learn a complex object recognition task from a few hundred labeled images. This is a small number compared with the apparent dimensionality of the problem (millions of pixels), but a child, or even a monkey, can apparently learn the same task from just a handful of examples. A comparison with real brains offers a second, probably related, challenge to learning theory: Classical learning theories of kernel machines correspond to one-layer architectures (see the "representer" theorem) and it seems that learning theory does not offer any general argument in favor of hierarchical learning machines for regression or classification. This is somewhat of a puzzle, since the organization of cortex, for instance the visual cortex, is strongly hierarchical. To add to the puzzle, hierarchical learning systems show superior performance in several engineering applications. The work conducted in the course of this project starts from the poverty of stimulus problem and assumes that it can be solved if the task to be learned is compatible with a hierarchical representation.

Motivated by the Air Force's interest in developing novel, powerful, adaptive learning techniques capable of achieving human-level performance on complex learning tasks, we have developed components of a mathematical theory of adaptive hierarchical learning architectures of the general type found in the visual cortex. The starting point for our preliminary mathematical work was a computational model we developed at MIT of the initial, feedforward flow of information in the primate

visual system for the task of recognizing images. This model is noteworthy in that it (1) agrees with a variety of physiological findings in different cortical areas of the ventral visual pathway, (2) is consistent with human psychophysics on rapid image categorization tasks, and (3) performs well on difficult recognition tasks compared to existing computer vision systems. The majority of the design decisions and engineering considerations were either wholly or in part motivated by the known neuroscience and physiology of visual cortex. Faithfulness to the biology is a hallmark of this model.

Our work to date has sought to abstract away from this model the elements we believe to be essential to understanding learning in hierarchies from an abstract, mathematical perspective. In particular, we have developed a mathematical description of an associated family of hierarchical architectures for learning comprised of a collection of definitions, lemmas and theorems which collectively highlight important and salient properties of such architectures. Most important among these properties is the notion of *invariance*. The theory we have developed characterizes *how and why* a hierarchical architecture can offer better generalization from few examples in terms capturing and exploiting *symmetries* in the physical world by way of learning invariances.

Intellectual Merit. Understanding how the brain works and reproducing its central capabilities in computers is arguably one of the greatest problems in science and engineering. This project directly contributes, from a mathematical point of view, to one of the major challenges for understanding intelligence: how does the brain represent and learn to recognize images and speech? How can we understand mathematically the hierarchical architectures that may underlie the solution of the poverty of stimulus problem?

Broader Impact. A mathematical framework to understand how to deal with the poverty of stimulus problem and how to discover and represent the complex invariances needed for robust object recognition - on the basis of few examples - will have a major impact in the fields of cognitive neuroscience and computer science, including artificial intelligence, robotics and surveillance systems. Some of these applications are discussed in Section 4.

2 Scientific Objectives

Our broad objective for this project was to develop theory and algorithms for hierarchical learning architectures of the general type found in the visual cortex. The specific goals we sought to address were as follows:

- Conduct a systematic study of the invariance properties of hierarchically organized models, including the role of templates and the number of layers in particular, and additionally further our understanding of the trade-off between invariance and selectivity.
- Explore the sample complexity of a classifier which uses the model as an unsupervised pre-processing step, as well as the sample complexity of learning a good hierarchical data-representation itself. A commonly held intuition is that the hierarchy, in the context of an unsupervised pre-processing step, induces a meaningful similarity measure that can be used to turn a difficult supervised classification problem into one that can be easily solved using existing, efficient off-the-shelf algorithms. The goal here is to ground this notion in a solid theoretical foundation, and understand the extent to which reliable generalization can be achieved with training samples smaller than what would otherwise be required without any pre-processing.
- Investigate automated, unsupervised learning of the templates: how to intelligently select the templates and/or directly learn the transformations (which may take the form of e.g., linear

operators) applied at each layer. Explore connections with (1) geometry of the data, and exploit manifold structure when appropriate via diffusion maps and other data-driven representations, (2) sparsity of the learned representations and the identification of salient features in the context of a hierarchy, and (3) optimization, convex analysis, and practical considerations essential to efficient implementations in applied settings.

- Explore "low-hanging" extensions of the basic hierarchical architecture: the effect of different pooling operations and architectural choices (e.g. the size of image analysis regions at each layer), possibly extending our work to encompass a larger class of architectures that have been previously proposed in the literature, as well as to suggest new ones. Interact with research involving notions of sparsity and data geometry in particular.

- Conduct detailed simulations involving images and video sequences with a hierarchical model for temporal data, and design an online algorithm allowing samples large enough to support the discovery of complex invariances in video. The intent of this objective is to highlight the usage of temporal dynamics to improve learning and real-time continuous adaptation in evolving visual environments.

3 Technical Approach

In terms of *analysis*, our general approach seeks to provide a theoretical and empirical characterization of generalization, invariance, and discrimination properties of abstract hierarchical models capturing the key organizational and computational themes of the visual cortex. Our approach to supporting *practical applications* involves the implementation of robust software tools for the design and simulation of flexible, massive-scale distributed hierarchical models.

3.1 Multi-Layer Hierarchies: A Plausible Computational Setting

The starting assumption guiding our work is that the sample complexity of (biological, feed-forward) object recognition is mostly due to geometric image transformations. Thus our main conjecture is that the computational goal of a feed-forward hierarchical model is to discount image transformations after learning them during "development". A complementary assumption is about the basic computational operation of such a model: we assume that (1) dot products between input vectors and stored templates (synaptic weights) are the basic operation; and (2) memory is stored in the synaptic weights through a Hebbian-like rule.

The models we consider are built from biologically plausible memory-based modules that learn transformations from unsupervised visual experience. The idea is that individual units can store during training "neural frames", that is, image patches of an object transforming for instance translating or looming. After training, the main operation consists of dot-products of the stored templates with a new image. The dot-products are followed by a transformation-averaging operation, which can be described as pooling. The main theorems we develop show that this 1-layer module provides (from a single image of any new object) a *signature* which is automatically invariant to global affine transformations and approximately invariant to other transformations. These results are derived in the case of random templates, using the Johnson-Lindenstrauss lemma in a special way; they are also valid in the case of sets of basis functions which are a frame. This one-layer architecture, though invariant, and optimal for clutter, is however not robust against local perturbations (unless a prohibitively large

set of templates is stored). A multi-layer hierarchical architecture is needed to achieve the dual goal of local and global invariance.

We show that a multi-layer hierarchical architecture of dot-product modules can learn in an unsupervised way geometric transformations of images and then achieve the dual goals of invariance to global affine transformations and of robustness to diffeomorphisms. A *linking conjecture* assumes that storage of transformed templates during development takes place via Hebbian-like developmental learning at the synapses in visual cortex. It follows that the cells' tuning will effectively converge during development to the top eigenvectors of the covariance of their inputs; the tuning of each cell is shown to converge to one of the eigenvectors. We assume that the development of this tuning takes place in stages, one layer at the time. We also assume that the development of tuning starts at the lowest layers with Gaussian apertures for the "simple" cells. Translations are effectively selected as the only learnable transformations during training by small apertures e.g. small receptive fields in the first layer. The solution of the associated eigenvalue problem predicts that the tuning of cells in the first layer identified with simple cells in V1 can be approximately described as oriented Gabor-like functions. This follows in a parameter-free way from properties of shifts, e.g. the translation *group*. Further, rather weak, assumptions about the spectrum of natural images imply that the eigenfunctions should in fact be Gabor-like with a finite wavelength which is proportional to to the variance of the Gaussian in the direction of the modulation. The theory also predicts an elliptic Gaussian envelope. Complex cells result from a local group average of simple cells. The hypothesis of a second stage of Hebbian learning at the level above the complex cells leads to wavelets-of-wavelets at higher layers representing local shifts in the cube of x,y, scale and orientation learned at the first layer. We derived simple properties of the number of eigenvectors and of the decay of eigenvalues as a function of the size of the receptive fields, to predict that the top learned eigenvectors and therefore the tuning of cells become increasingly complex and closer to each other in eigenvalue.

3.2 Approach to the Data Representation Problem

In much of the analysis described above, templates connecting the model to a particular problem were learned and stored in some form at each layer on the basis of accumulated experience. We studied the problem of learning data representations in an abstract statistical setting in an effort to better understand how learning at each layer can proceed efficiently.

Our approach follows from the observation that many data representation algorithms (e.g., clustering, sparse-coding) are but instances of a general problem differentiated only by the fact that the empirical problem solved by each algorithm is defined only by a set of algorithm specific constraints. A central question is then: *Is there a common objective underlying the learning process of different data representation algorithms?* This question, although fundamental, was essentially unexplored. Answering this question has been key to understanding and improving the models proposed thus far, and for guiding the development of new ones. In particular, this work has suggested new computational routines to adaptively tune the parameters in the representation learning algorithms to be maximally effective and robust to noise. More generally, the theoretical results we have explored seek to predict the benefits of different algorithmic solutions: Is there an advantage in stacking multiple learning layers? What is the impact of potentially nonconvex problems?

The above questions were tackled within a statistical learning framework. We assume a training set X_n to be a sample from a distribution ρ on a Hilbert space[1] $\mathcal{X}$, with a focus on the setting where

[1]This assumption allows to model visual/audio signals, more general setting (metric space) spaces may be investigated

ρ is supported on, or close to, a d-dimensional manifold[2] $\mathcal{M}$. In contrast to classical studies in signal processing, in this learning setting the emphasis is on the data generating distribution, rather than signals from some class.

4 Progress Made & Results Obtained

An efficient software implementation of the particular hierarchical model described in [18] was completed in year 2009-10. Following the complexity analysis in this reference, the implementation achieves complexity linear in the number of model layers by reorganizing the computations to avoid redundant computation. A comprehensive, high-throughput GPU implementation was begun in year 2009-10, and has since been completed [15]. This freely-available software package is general, and captures not only the model described in [18] but also several other families of deep hierarchical models, including convolutional neural networks built from multiple pooling, sampling, normalization and filtering possibilities. The package provides a selection of several training algorithms, supporting both batch and online approaches. This investment in software and algorithmic infrastructure has allowed for large-scale empirical experimentation and validation, and provides a rapid platform with which to test hypotheses. The last (supervised) layer of the architecture was subsequently developed into a multi-category classification toolbox which utilizes the underlying distributed infrastructure. Extensive experimental analyses have been done using these tools on state of the art computer vision datasets, including large-scale datasets.

Thorough empirical simulations exploring invariance and discrimination properties of multi-layer hierarchical representations were carried out. We tested the impact of a rotation invariant initial kernel (based on image histograms), and compared to the baseline performance of non-invariant initial kernels [1, 19]. The induced invariance properties were analyzed in the context of image classification experiments. Invariance of the initial kernel did lead to substantial invariance in the top-level derived kernel, as seen empirically and predicted theoretically. We also conducted further simulations confirming that the hierarchy – and its corresponding invariance to transformations – can indeed reduce the sample complexity of a supervised image discrimination task, when used as a representational pre-processing step [19].

Encouraged by the these experimental results, a generalized architecture allowing for arbitrary pooling functions and more general layer-wise feature maps was developed theoretically [2] and tested empirically. We explored architectures involving max, max-absolute-value, and average pooling, as well as architectures employing Kernel PCA (KPCA) embeddings at each layer [19]. In these experiments, the embeddings were derived from samples of patches extracted from natural images. We found that substantial computational savings can be realized by taking just a few dominant KPCA components instead of storing the original templates (patches) at each layer of the hierarchy. This empirical investigation constituted a first step towards, and motivation for, a body of work in subsequent years of the project (described below) investigating unsupervised template learning possibilities which exploit the geometry of image patches presented to a model.

Experiments addressing video sequences were also performed (see [16], and figures therein). These simulations compared localized PCA-based representations learned at each layer of a hierarchy given static natural images versus evolving video input. It was found that motion is often important

to deal with other kinds of data, e.g. graphs or probability measures.

[2] A "manifold plus noise" model is useful to model high dimensional data and covers the special classic case $\mathcal{X} = \mathbb{R}^p$ and ρ nonsingular and absolutely continuous.

for learning, and learning models from video can give improved performance over training with static images. Finally, we additionally considered a modified hierarchical model which accounts for peripheral vision effects [8]. A GPU implementation of this model was completed, and a module for distributed computing has been designed and implemented.

In response to the theoretical goals of this project, we have developed a theory of the ventral stream in visual cortex which predicts the emergence of a hierarchical architecture from assuming that the computational goal of the ventral stream is to learn transformations during development and to become invariant to such transformations [16]. This aspect of the project is described in more detail below in Section 4.1. At higher representation levels, the nature of the transformations applied to data flowing through the hierarchy is complex, and we have started to develop a regularization theory of data representation in learning. This theory shows that several seemingly diverse algorithms are in fact optimizing the same generalization error functional under different prior assumptions, and offers a unifying perspective within which one can design and tune learning algorithms. Further detail is provided in Section 4.2.

Finally, an extension of the model described in [18] was developed and applied to strings describing amino acid sequences, towards the ultimate goal of understanding and predicting the folding of proteins [17]. This application of hierarchical models to problems in immunology involved the design and implementation of new kernels and string-optimized architectures (e.g. pooling operations and layering considerations), and led to state of the art performance on a set of benchmark challenges. We describe this set of results in Section 4.3.

4.1 Invariance to Transformation in Hierarchical Architectures: A Theory of the Ventral Steam

Despite significant advances in sensory neuroscience over the last five decades, a true understanding of the basic functions of the ventral stream in visual cortex has proven to be elusive. The theory we have developed [16] proposes that the main computational goal of the ventral stream is to provide a representation of new objects/images which is invariant to geometric transformations learned during development, is stable with respect to deformations and noise and is sufficiently discriminative to access higher level memory and categorization stages. Invariant representations considerably reduce the sample complexity of learning, allowing recognition of new object classes from very few examples - in the limit just one.

There are two main sets of results, one focusing on properties of the basic simple-complex cell modules and a second focusing on multi-layer architectures composed of such modules. The first asserts that the probability distribution induced by the action of a group is a unique and invariant property of the image. Nonlinear sigmoidal functions of the empirical estimate of one dimensional projections of the distribution are proven to be a unique, invariant signature associated with the image. Surprisingly, complex cells in V1 can naturally implement such estimates. We further prove that a hierarchical architecture provides the required properties of invariance, stability and discriminability. According to the second set of theorems, a hierarchical architecture, in addition to the desired properties of invariance, stability and discriminability, is also matched to the hierarchical structure of the visual world and to the need to retrieve items from memory at various levels of size and complexity. Our work outlines a theoretical foundation that should readily apply to several existing hierarchical architectures, such as HMAX, convolutional networks and related feed-forward models of the visual system, formally characterizing their properties and explaining recent successes of hierarchical architectures of the convolutional type on several recent visual and speech recognition tests. Moreover, a

very similar analysis should apply to other sensory modalities, such as speech recognition, given that these modalities are also strongly hierarchical in nature.

Another key result of this theory is that a hierarchical architecture of the modules introduced in Section 3 with "receptive fields" of increasing size, provides global invariance and stability to local perturbations (and in particular tolerance to local deformations). Interestingly, the whole-parts theorem implicitly defines "object parts" as small patches of the image which are locally invariant and occur often in images. The theory predicts a stratification of ranges of invariance: size and position invariance should develop in a sequential order meaning that smaller transformations are invariant before larger ones, in earlier layers of the hierarchy.

4.2 A Statistical Framework for Learning Data Representations

A fundamentally important component of any hierarchical model is the data representation chosen at each layer. Typical choices include sparse-coding schemes or K-means representations of previously viewed input patches. We have (1) systematically studied the theoretical, computational and empirical properties of different learning/coding schemes [5, 4], and (2) developed a regularization framework for learning and exploiting the hierarchical structure describing the statistical relations between large numbers of categories [13, 14]. The framework is a synthesis of unsupervised and supervised learning and uses concepts from graphical models and vector valued function approximation. Algorithms stemming from the theoretical studies described above have been implemented, tested and optimized.

The approach to learning data representations we considered assumes noiseless data. Given this assumption, we proceeded with two complementary approaches. In the first approach, the quality of a representation is determined by the distribution obtained by transforming the data. We started investigating this perspective in [5] showing that *all* data representation algorithms in a broad class capturing many popular approaches estimate the data generating distribution with respect to the W_2 Wasserstein metric. Concentration results for the empirical distribution are key in this setting and are an active field of study in Optimal Transport Theory. The second approach is more geometric and is based on concepts from regularization theory of ill-posed variational problems. Using topological constraints, we defined a suitable notion of Moore-Penrose generalized solutions for expected reconstruction error minimization. We showed that many data representation algorithms define a regularized solution to this problem. In particular, we studied piece-wise algorithms, such as K-flats in [4]. These methods return sparse representations, similar to popular sparse coding methods, but are considerably more amenable to theoretical analyses.

4.3 Towards a Mathematical Foundation of Immunology: Hierarchical Models for Amino Acid Chains

This set of contributions [17] attempts to set a mathematical foundation of immunology and amino acid chains. To measure the similarities of these chains, a kernel on strings was defined using only the sequence of the chains and a good amino acid substitution matrix (e.g. BLOSUM62). This kernel was then used in learning machines to predict binding affinities of peptides to human leukocyte antigens DR (HLA-DR) molecules. On both fixed allele (Nielsen and Lund 2009) and pan-allele (Nielsen et.al. 2010) benchmark databases, our algorithm achieves state-of-the-art performance. The kernel is also used to define a distance on an HLA-DR allele set based on which a clustering analysis precisely recovers the stereotype classifications assigned by the WHO (Nielsen and Lund 2009, and Marsh et.al. 2010). These results suggested that the kernel we have designed relates well the chain structure

of both peptides and HLA-DR molecules to their biological functions, and suggested that it offers a simple but powerful and promising methodology to immunology and amino acid chain studies.

5 Significance of Results & Impact on Science

Three outstanding questions have continued to puzzle researchers in both neuroscience and the computational/mathematical sciences: (1) Why do deep hierarchical models, versus shallow models, work as well as they do? (2) How does a model's architecture impact performance and why? (3) How can a model be efficiently trained or designed to reduce training time and/or sample complexity, given a particular task? Our work has sought to address these questions from both theoretical and empirical perspectives. The significance and impact of our contributions may be summarized as follows:

- The theoretical analyses explain in rigorous terms how a model captures and achieves invariance to transformations, and shows that this invariance is what is responsible for good performance. This work suggests how a model might be designed to achieve greater accuracy at lower sample complexities.
- A surprising implication of the theoretical results we have obtained is that the computational goals and several of the tuning properties of cells in the ventral stream may follow from symmetry properties (in the sense of physics) of the visual world through a process of unsupervised correlational learning, based on Hebbian synapses.
- The theory of the ventral stream developed in [16] may be considered to be perhaps the first attempt at a theory of learning in visual cortex since the theory of Hubel and Wiesel.
- Our empirical simulations validate the theory, and provide a better understanding as to the role and influence of architectural design choices upon performance and discrimination properties.
- The study of template learning we have undertaken highlights connections with and provides insights into developmental learning (ongoing work with S. Ullman, Weizmann Institute of Science).
- The present work has led to contact with AFLR (Todd Howlett and Yurij Luzanov) in the context of a project modeling feedback in vision systems (funded – to be carried out mainly by G. Kreiman (Harvard)). The study of dynamic feedback hierarchies is a natural extension of the work we have undertaken here, and may be informed by the results we have reported.
- The software tools we have developed provide a comprehensive, easy-to-use platform that can be used to analyze massive datasets with deep hierarchical models. The distributed GPU functionality provides immediate cost savings, and drastically reduces the test-debug cycle time. Our tools have found success in real applications: the lab of Sebastian Seung (MIT) is currently using our software to train massive convolutional networks (millions of weights) for segmentation of neuronal synapses from 3-D two-photon microscopy image stacks.
- A goal of this project has been to establish a statistical theory of the learning of data representations that occurs at each layer of a hierarchical model, akin to the classic one for supervised learning. The latter has provided an understanding of the properties of popular supervised algorithms and suggested a variety of generalizations and new algorithms. Indeed, the close interplay of theory and practice is often cited as one of the main reasons why supervised machine learning has made so much progress over the years. In our work, we developed a statistical

theory of learning data representations within a multidisciplinary approach drawing computational and mathematical tools from Computer Science, Engineering, and Mathematics [4, 5]. As in supervised learning, this new theory provides new understanding, and more adaptive and efficient data representation algorithms. In particular, it has guided the design of new invariant, multiscale algorithms for learning under weak supervision. Our results show, for example, that there is an optimal dictionary (codebook) size for a given template representation problem in the sense that it optimizes a bias-variance tradeoff. (Note, this is not equal to the number of datapoints, and can be much smaller.) This result validates in precise terms the perspective that a good representation is one that provides *compression*.

- The development of hierarchical models for studying amino acid chains has the potential to offer substantial, concrete contributions to our understanding of disease and the design of effective vaccines. Large scientific and industrial enterprises are engaged in efforts to produce new vaccines from synthetic peptides. The study of peptide binding to appropriate alleles is a major part of our effort, and we support the use of string kernels for peptide binding prediction as well as for the classification of supertypes of the major histocompatibility complex (MHC) alleles.

Publications Resulting From Research

[1] J. Bouvrie, T. Poggio, L. Rosasco, S. Smale, A. Wibisono, "Generalization and Properties of the Neural Response", MIT CSAIL Tech Report 2010-051/CBCL-292, Massachusetts Institute of Technology, Cambridge, MA, November 19, 2010.

[2] J. Bouvrie, L. Rosasco, and T. Poggio. "On Invariance in Hierarchical Models". Advances in Neural Information Processing Systems (NIPS) 22, 2009.

[3] J. Bouvrie, L. Rosasco, G. Shakhnarovich, and S. Smale. "On the Shannon Entropy of the Neural Response", MIT CSAIL Tech Report 2009-049/CBCL Paper 281, Massachusetts Institute of Technology, Cambridge, MA, October, 2009.

[4] G.D. Canas, T. Poggio, L. Rosasco. "Learning Manifolds with K-Means and K-Flats", Advances in Neural Information Processing Systems (NIPS) 25, 2012.

[5] G.D. Canas, L. Rosasco. "Learning Probability Measures with respect to Optimal Transport Metrics", Advances in Neural Information Processing Systems (NIPS) 25, 2012.

[6] S. Chikkerur and T. Poggio, "Approximations in the HMAX Model", MIT-CSAIL-TR-2011-021/CBCL-298, Massachusetts Institute of Technology, Cambridge, MA, April 14, 2011.

[7] L. Isik, J.Z. Leibo and T. Poggio. "Learning and disrupting invariance in visual recognition with a temporal association rule", Front. Comput. Neurosci. 6:37. doi: 10.3389/fncom.2012.00037, June 25, 2012.

[8] L. Isik, J.Z. Leibo, J. Mutch, S.W. Lee, and T Poggio. "A hierarchical model of peripheral vision", MIT-CSAIL-TR-2011-031/CBCL-300, Massachusetts Institute of Technology, Cambridge, MA, June 2011.

[9] T. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre "HMDB: A Large Video Database for Human Motion Recognition," ICCV 2011.

[10] J.Z. Leibo, J. Mutch, T Poggio. "How can cells in the anterior medial face patch be viewpoint invariant?" Presented at COSYNE 2011, Salt Lake City, UT. Available from Nature Precedings at dx.doi.org/10.1038/npre.2011.5845.1, 2011.

[11] J.Z. Leibo, J. Mutch, T Poggio. "Learning to discount transformations as the computational goal of visual cortex" Presented at FGVC/CVPR 2011, Colorado Springs, CO. Available from Nature Precedings at dx.doi.org/10.1038/npre.2011.6078.1, 2011.

[12] J.Z. Leibo, J. Mutch, T. Poggio, "Why The Brain Separates Face Recognition From Object Recognition", Advances in Neural Information Processing Systems (NIPS) 24, 2011.

[13] Y. Mroueh, T. Poggio, L. Rosasco and J.J. Slotine. "Multi-class Learning with Simplex Coding", In Advances in Neural Information Processing Systems, NIPS 2012.

[14] Y. Mroueh, T. Poggio and L. Rosasco. "Regularization Predicts While Discovering Taxonomy", First Workshop on Fine-Grained Visual Categorization at CVPR, 2011.

[15] J. Mutch, U. Knoblich and T. Poggio, "CNS: a GPU-based framework for simulating cortically-organized networks". MIT-CSAIL-TR-2010-013 / CBCL-286, Massachusetts Institute of Technology, Cambridge, MA, February 26, 2010.

[16] T. Poggio, J. Mutch, F. Anselmi, L. Rosasco, J.Z. Leibo, and A. Tacchetti, "The computational magic of the ventral stream: sketch of a theory (and why some deep architectures work)". MIT-CSAIL-TR-2012-035, Massachusetts Institute of Technology, Cambridge, MA, December 29, 2012.

[17] Wen-Jun Shen, Hau-San Wong, Quan-Wu Xiao, Xin Guo, and Stephen Smale. "Towards a Mathematical Foundation of Immunology and Amino Acid Chains", `arXiv:1205.6031`, June 25, 2012.

[18] S. Smale, L. Rosasco, J. Bouvrie, A. Caponnetto, and T. Poggio. "Mathematics of the Neural Response", Foundations of ComputationalMathematics, Vol. 10(1), pp.67–91, Feb 2010.

[19] A. Wibisono, J. Bouvrie, L. Rosasco and T. Poggio. "Learning and Invariance in a Family of Hierarchical Kernels", MIT CSAIL Tech Report MIT-CSAIL-TR-2010-035/CBCL Paper 290, Massachusetts Institute of Technology, Cambridge, MA, July, 2010.